Dataset
Comparison vs adenocarcinomas and squamous cell carcinomas
https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3627
Pipeline Steps:

1. The student should first test for outlier samples and provide visual proof.
    a. Remove these outliers
2. Then, filter out genes that have low expression values using some criterion (e.g. average gene expression < 50 for Affy MAS4/MAS5 data)
3. Conduct some method of feature selection with a statistical test or other machine learning method. The type of test will depend upon how many factor levels are included in your data set. For example. two conditions would require a two-sample test, while greater than two conditions would require other tests
4. Provide the number of genes retained with the associated score (p-value, weight, test statistic, etc.) and threshold value that you used
5. Plot the scores of those genes retained in a histogram
6. Next, subset your data by the genes that you determined and use one of the clustering or dimensionality reduction methods discussed in class to visualize the samples in two-dimensional space (xy scatter plot, dendrogram, etc.).
7. Using these linear projections of the original data (i.e. cluster centroids, principal components, latent variables, etc.), use a classification method to classify the samples into their respective classes. Make sure to color the samples appropriately by their predicted class membership and use different symbols for the actual class memberships
8. Finally, using the top 5 discriminant genes (positive and negative direction) from your analysis, go to NCBI's DAVID (http://david.abcc.ncifcrf.gov/home.jsp) and look up the gene information. Provide the gene name and functional information (associated pathways, GO terms, etc) for these 10 genes.

Code

1. **The student should first test for outlier samples and provide visual proof.**
    a. **Remove these outliers**

```
> library(GEOquery)
> gse <- getGEO("GSE10245", GSEMatrix = TRUE, getGPL=FALSE)

# retrieve expression data from GSE
> dat <- exprs(gse[[1]])

#create the annotation vector classifying each sample as either Adeno or Squamous
> ann <- gse[["GSE10245_series_matrix.txt.gz"]]@phenoData@data[["disease state:ch1"]]
```

Correlation heatmap, Clustering dendrogram, PCA plot, CV vs. mean plot, Avg correlation plot all can be used to identify the outliers

```
> library(gplots)
```

CV vs. mean plot

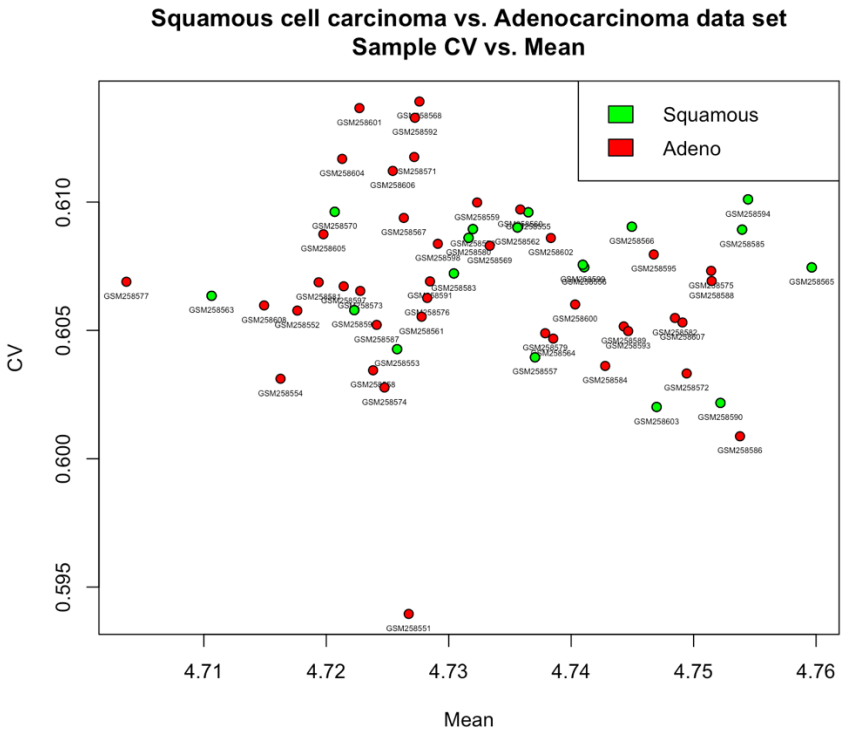
```
# calculate mean for each sample, removing NA values
> dat.mean <- apply((dat),2,mean, na.rm=TRUE)

# calculate std for each sample, removing NA values
> dat.sd <- sqrt(apply((dat),2,var, na.rm=TRUE))

#calculate cv
> dat.cv <- dat.sd/dat.mean
```

#plot CV vs. mean and color each sample group differently
> plot(dat.mean,dat.cv,main="Squamous cell carcinoma vs. Adenocarcinoma data set\nSample CV vs. Mean",xlab="Mean",ylab="CV",col='blue',cex=1.5,type="n")

 # find the columns that correlate to scc
> scc <- which(ann == "squamous cell carcinoma")
> adn <- which(ann == "adenocarcinoma")

#plot each sample group a different color to look for outliers
> points(dat.mean[scc],dat.cv[scc],bg="green",col=1,pch=21)
> points(dat.mean[adn],dat.cv[adn],bg="red",col=1,pch=21)
> text(dat.mean, dat.cv, label = dimnames(dat)[[2]], pos=1, cex=0.4)

> legend("topright", legend=c("Squamous", "Adeno"), fill=c("green", "red"), cex = 1)



CV vs. Mean Plot Takeaways

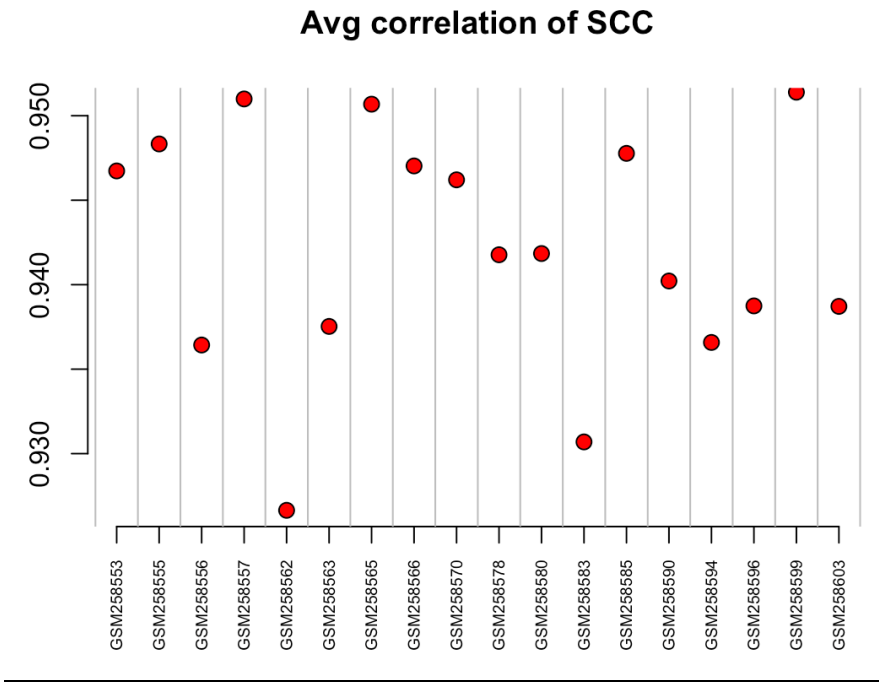
The major outlier appears to be GSM258551 for the Adenocarcinoma

Average Correlation Plot

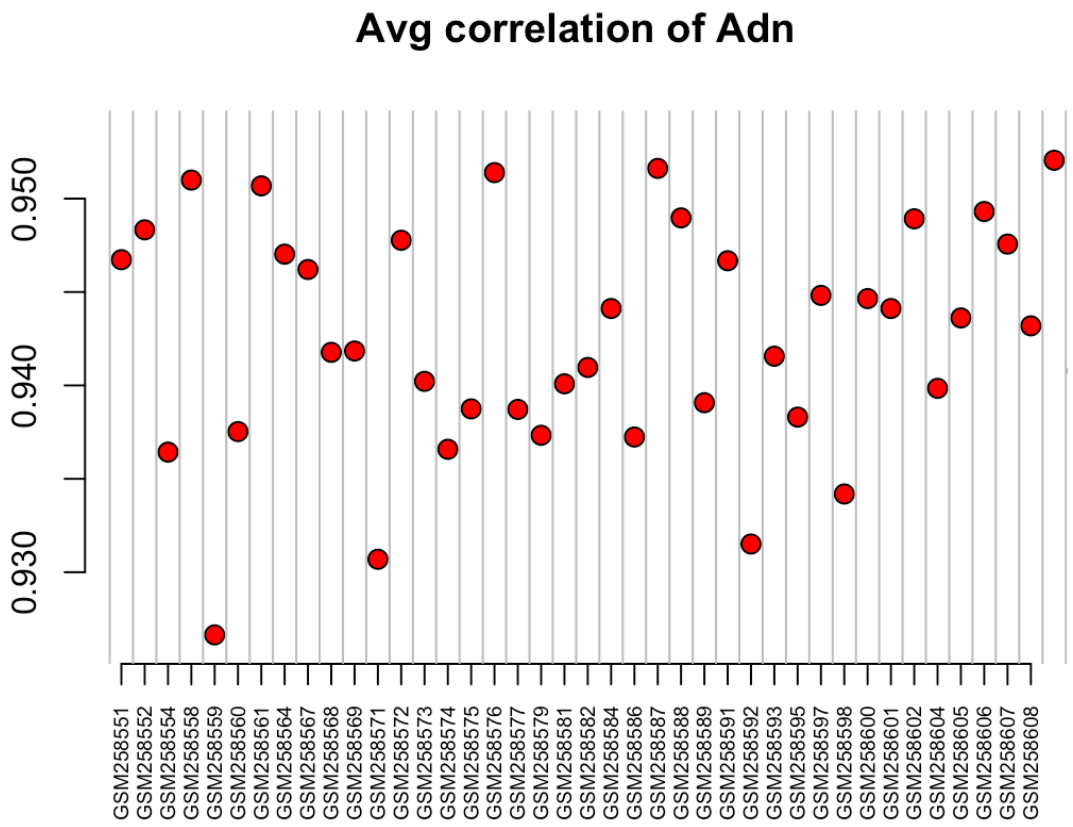
 # get correlation of data
> dat.cor<-(cor((dat), use="pairwise.complete.obs"))
> dat.avg <- apply(dat.cor, 1, mean)

 # split average vector into just the averages for each sample sets
> dat.avg.scc <- dat.avg[scc]
> dat.avg.adn <- dat.avg[adn]

 # Plot average correlation plot for each sample set
> par(oma=c(3,0.1,0.1,0.1))
> plot(c(1,length(dat.avg.scc)),range(dat.avg.scc),type="n",xlab="",ylab="Avg r",main="Avg correlation of SCC",axes=F)
> points(dat.avg,bg="red",col=1,pch=21,cex=1.25) #put points down on graph
>
axis(1,at=c(1:length(dat.avg.scc)),labels=names(dat.avg.scc),las=2,cex.lab=0.4,cex.axis=0.6) #label the x axis with the sample names
> axis(2)
> abline(v=seq(0.5,62.5,1),col="grey") #add the gray bars to line up samples with their avg correlation



```
> par(oma=c(3,0.1,0.1,0.1))
> plot(c(1,length(dat.avg.adn)),range(dat.avg.adn),type="n",xlab="",ylab="Avg
r",main="Avg correlation of Adn",axes=F)
> points(dat.avg,bg="red",col=1,pch=21,cex=1.25) #put points down on graph
>
axis(1,at=c(1:length(dat.avg.adn)),labels=names(dat.avg.adn),las=2,cex.lab=0.4,cex.
axis=0.6) #label the x axis with the sample names
> axis(2)
> abline(v=seq(0.5,62.5,1),col="grey") #add the gray bars to line up samples with
their avg correlation
```



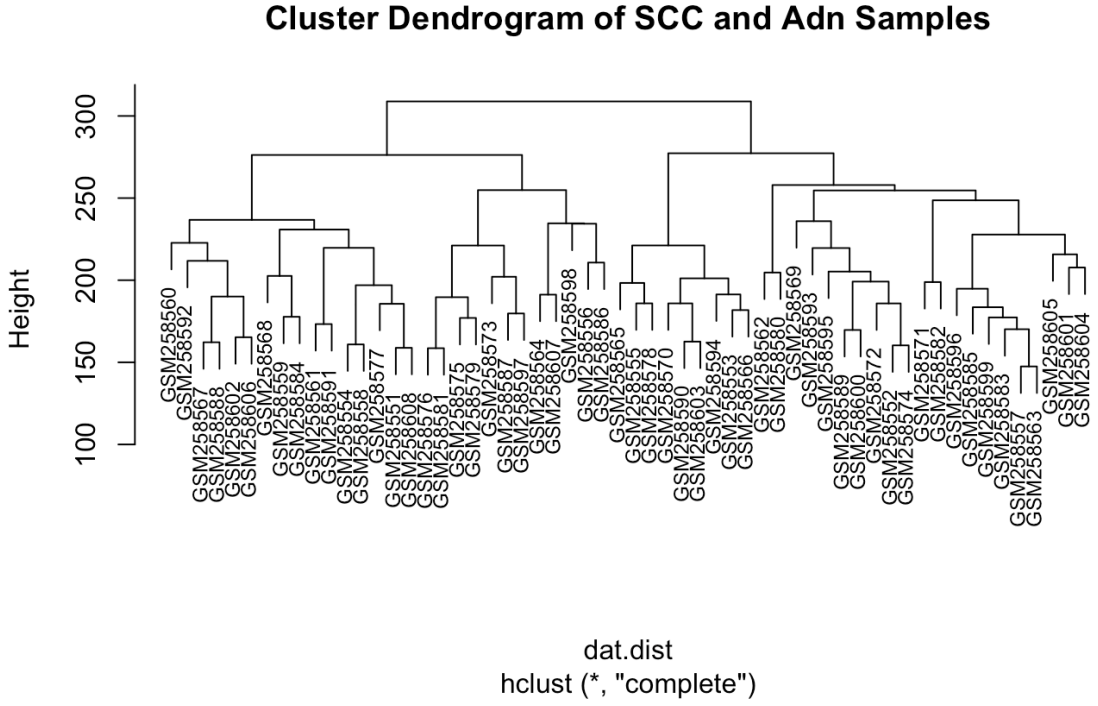
<u>Avg Correlation Plot Takeaways</u>

For the SCC samples, the average correlation of GSM258562 and GSM258583 appear to be outliers and for the Adn samples, GSM258559 appears to be an outlier. These samples do not correlate with the samples believed to be outliers from the CV vs. mean plot.

Hierarchical Clustering Dendrogram

 # get pairwise distance
> dat.dist <- dist(t(dat), method="euclidean")

 # plot hierarchical tree
> dat.clust <- hclust(dat.dist)
> plot(dat.clust,labels =row.names(dat), cex=0.75, main = "Cluster Dendrogram of SCC and Adn Samples")



Hierarchical Clustering Dendrogram Plot Takeaways:

There appears to be no outstanding outlier samples, and because the potential outliers do not correlate from the CV vs. mean plot and the Avg. correlation plot, I do not believe there is a need to remove any samples from this data set.

2. **Then, filter out genes that have low expression values using some criterion (e.g. average gene expression < 50 for Affy MAS4/MAS5 data)**

```
# find mean of gene count for whole data set
> mean(dat)
[1] 4.733484
> 4.7 * .25
[1] 1.175

# find all genes that have mean expression less than 25% of the mean
> row.means <- rowMeans(dat)
> remove <- which(row.means < 1.175)

> remove
  1552411_at 1553039_a_at   1566956_at
         114          557         8529
  1569415_at  214391_x_at   38707_r_at
        9206        23691        54320
    71933_at
       54587
7 genes to be removed

# remove these genes
> dat.1 <- dat[-remove,]
```

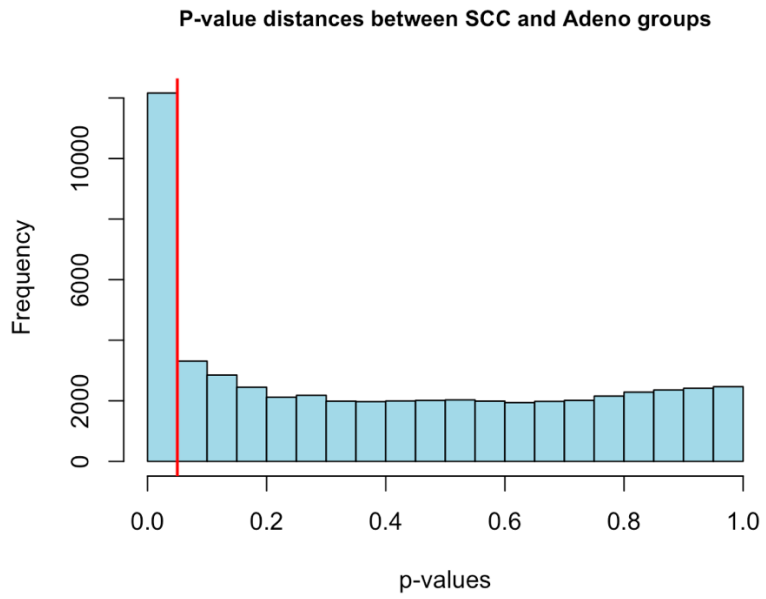
3. Conduct some method of feature selection with a statistical test or other machine learning method. The type of test will depend upon how many factor levels are included in your data set. For example. two conditions would require a two-sample test, while greater than two conditions would require other tests

2 Sample T-test

```
# Student's t-test
t.test.all.genes <- function(x,s1,s2) {
x1 <- x[s1]
x2 <- x[s2]
x1 <- as.numeric(x1)
x2 <- as.numeric(x2)
t.out <- t.test(x1,x2, alternative="two.sided",var.equal=T)
out <- as.numeric(t.out$p.value)
return(out)
}

# subscript scc vs adn
> scc <- which(ann == "squamous cell carcinoma")
> adn <- which(ann == "adenocarcinoma")

#t-test on all genes
> pv <- apply(dat.1, 1, t.test.all.genes, s1=scc, s2=adn)
```

4. Provide the number of genes retained with the associated score (p-value, weight, test statistic, etc.) and threshold value that you used

#graph p-values on a histogram
> hist(pv, col="lightblue", xlab="p-values", main="P-value distances between SCC and Adeno groups", cex.main=0.9)
> abline(v=.05,col=2,lwd=2)



#Get all values that are less than p of .05
> pv.05 <- pv < 0.05
> sum(pv.05)
1] 12161 p-values <0.05

#values of p < 0.01
> pv.01 <- pv < 0.01
> sum(pv.01)
[1] 7416 p-values < 0.01

#probeset threshold = 0.05/total probesets
> threshold <- 0.05/54675
[1] 9.144947e-07

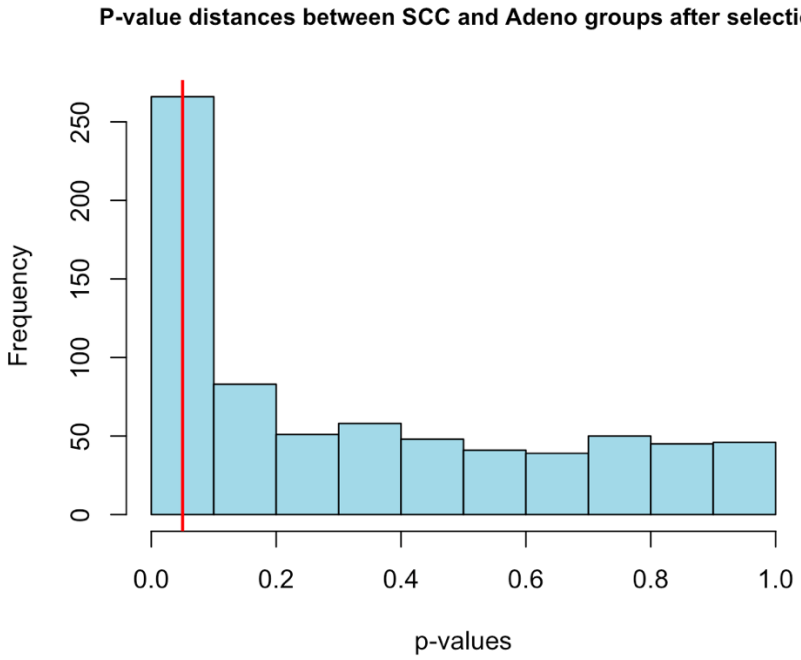
#how many p-values below this number
> pv.threshold <- pv < threshold
> sum(pv.threshold)
[1] 727 p-values below the threshold.

5. Plot the scores of those genes retained in a histogram

 # get a vector of the indexes where the pv is under the threshold
> pv.thresh.loc <- which(pv < threshold)

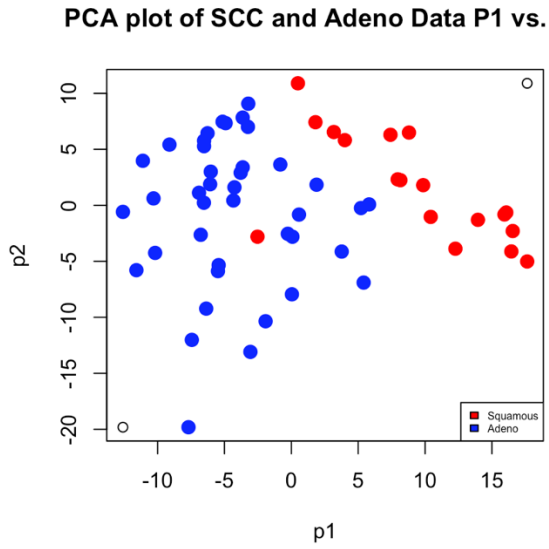
 # select these pv's from the original pv value
> pv.final <- pv[pv.thresh.loc]

 # Histogram plot of retained genes
> hist(pv.final, col="lightblue", xlab="p-values", main="P-value distances between SCC and Adeno groups after selection", cex.main=0.9)
> abline(v=.05,col=2,lwd=2)

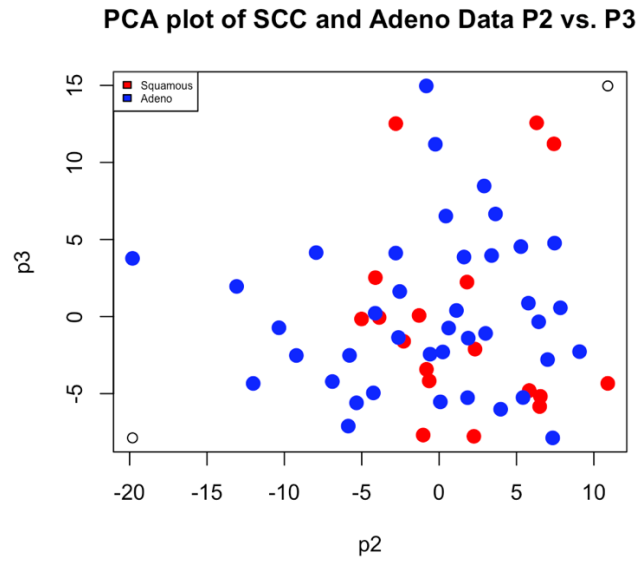


6. Next, subset your data by the genes that you determined and use one of the clustering or dimensionality reduction methods discussed in class to visualize the samples in two-dimensional space (xy scatter plot, dendrogram, etc.).

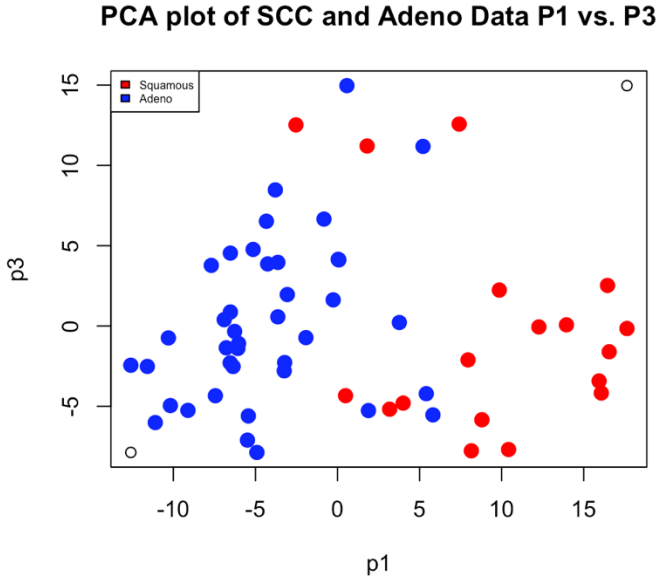
```
# subset data by genes retained from part 5
> include <- names(pv.final)
> dat.2 <- dat.1[include,]

> dat.pca <- prcomp(t(dat.2),cor=F)
> dat.loadings <- dat.pca$x[,1:3]
>
plot(range(dat.loadings[,1]),range(dat.loadings[,2]),xlab='p1',ylab='p2',main=
'PCA plot of SCC and Adeno Data P1 vs. P2')
> points(dat.loadings[,1][scc],
dat.loadings[,2][scc],col='red',pch=16,cex=1.5)
> points(dat.loadings[,1][adn],
dat.loadings[,2][adn],col='blue',pch=16,cex=1.5)
> legend("bottomright", legend=c("Squamous", "Adeno"), fill=c("red",
"blue"), cex = 0.5)
```




```
>
plot(range(dat.loadings[,2]),range(dat.loadings[,3]),xlab='p2',ylab='p3',main
='PCA plot of SCC and Adeno Data P2 vs. P3')
> points(dat.loadings[,2][scc], dat.loadings[,3][scc],col='red',pch=16,cex=1.5)
> points(dat.loadings[,2][adn],
dat.loadings[,3][adn],col='blue',pch=16,cex=1.5)
> legend("topleft", legend=c("Squamous", "Adeno"), fill=c("red", "blue"),
cex = 0.5)
```

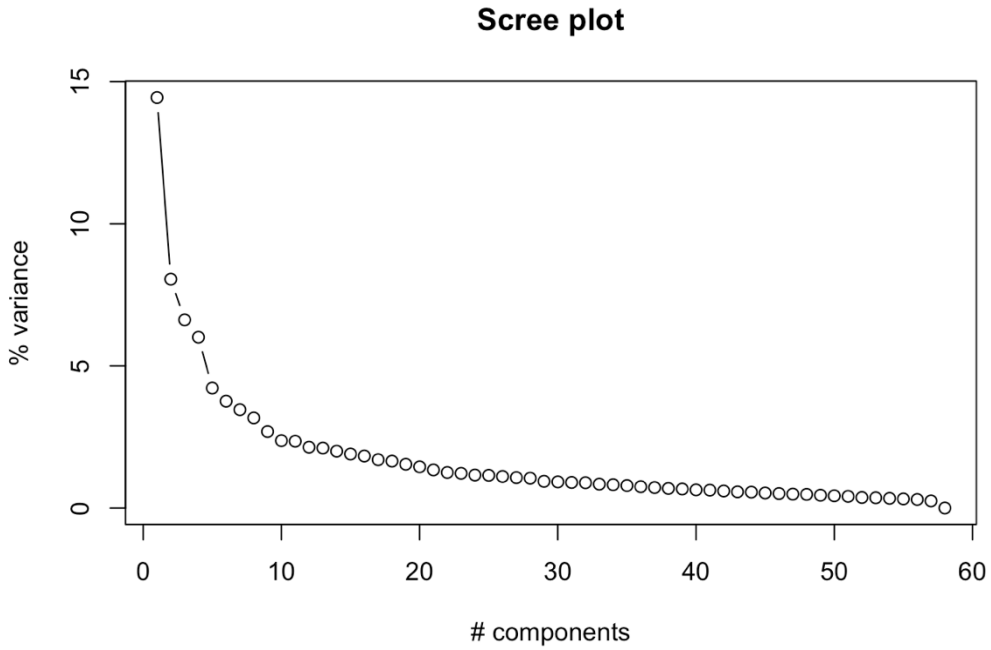


>
plot(range(dat.loadings[,1]),range(dat.loadings[,3]),xlab='p1',ylab='p3',main='PCA plot of SCC and Adeno Data P1 vs. P3')
> points(dat.loadings[,1][scc], dat.loadings[,3][scc],col='red',pch=16,cex=1.5)
> points(dat.loadings[,1][adn], dat.loadings[,3][adn],col='blue',pch=16,cex=1.5)
> legend("topleft", legend=c("Squamous", "Adeno"), fill=c("red", "blue"), cex = 0.5)



Scree Plot of eigenvalues:

```
> dat.pca.var <- round(dat.pca$sdev^2 / sum(dat.pca$sdev^2)*100,2)
> plot(c(1:length(dat.pca.var)),dat.pca.var,type='b',xlab='#
components',ylab='% variance',main='Scree plot')
```



Scree plot takeaways: First 4 eigen values are all representative of over at least 5% of the variance found in the data. Elbow appears to be around PC8.

7. Using these linear projections of the original data (i.e. cluster centroids, principal components, latent variables, etc.), use a classification method to classify the samples into their respective classes. Make sure to color the samples appropriately by their predicted class membership and use different symbols for the actual class memberships

```
> library(MASS)

# get all PCs into a single data matrix
> dat.pca.all <- dat.pca$x

# add classification column at the end
> dat.pca.all <- data.frame(dat.pca.all, ann)

# select training set
> scc
 [1]  3  5  6  7 12 13 15 16 20 28 30 33
[13] 35 40 44 46 49 53

# select the first 9 scc (50%) samples for training
> scc.training <- scc[1:9]

> adn
 [1]  1  2  4  8  9 10 11 14 17 18 19 21
[13] 22 23 24 25 26 27 29 31 32 34 36 37
[25] 38 39 41 42 43 45 47 48 50 51 52 54
[37] 55 56 57 58

# select the first 21 adn samples (50%) for training
> adn.training <- adn[1:21]

# create new matrix of the training PCA data
> train <- dat.pca.all[c(scc.training, adn.training),]

# create labels for training matrix
> train.lab <- train[,59]

# remove annotations from the training set
> train <- train[,c(1:58)]
```

 # select only the first 8 PCs to run LDA on as the rest do not count for much variance in the data set.

```
> model <- lda(train.lab~., train[,c(1:8)])
> out <- predict(model, train[,c(1:8)])
> table(out$class,train.lab)
                          train.lab
                           adenocarcinoma
  adenocarcinoma                       21
  squamous cell carcinoma               0
                          train.lab
                           squamous cell carcinoma
  adenocarcinoma                                 2
  squamous cell carcinoma                        7
```

2 mis-classified SCC samples that have been classified as Adn instead.

 # apply this model to the entire data set

```
> out <- predict(model, dat.pca.all[,c(1:58)])
> table(out$class,ann)
                          ann
                           adenocarcinoma
  adenocarcinoma                       37
  squamous cell carcinoma               3
                          ann
                           squamous cell carcinoma
  adenocarcinoma                                 2
  squamous cell carcinoma                       16
```

5 total mis-classified samples. 3 Mis-classified Adn samples and 2 mis-classified SCC samples.

 # Plot samples with true class and predicted class

```
> predicted.classes <- out$class
```

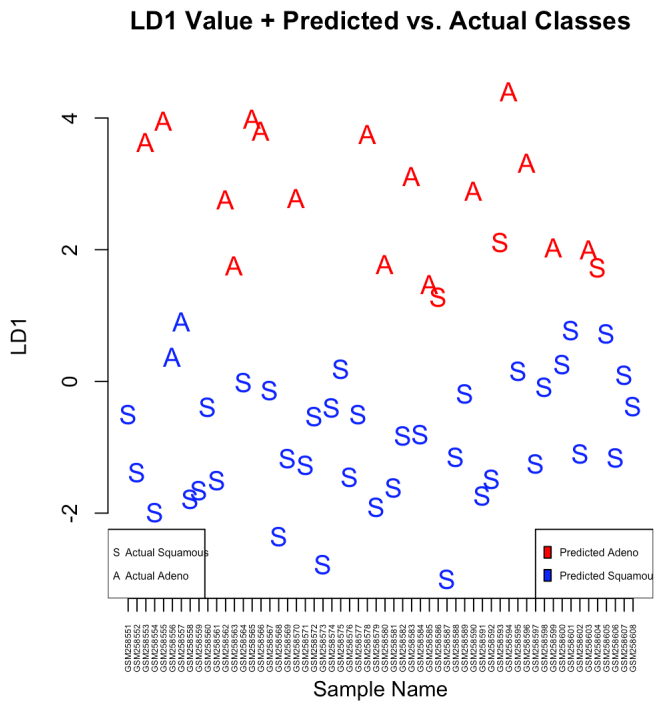
 # combine predicted class vector and out$x values

```
> predicted.classes <- cbind(predicted.classes, out$x)
```

 # plot matrix of LD1 with predicted and real class
> plot(c(1,length(out$x)),range(out$x),type="n",xlab="Sample Name",ylab="LD1",main="LD1 Value + Predicted vs. Actual Classes",axes=F)
>
axis(1,at=c(1:length(out$x)),labels=row.names(dat.pca.all),las=2,cex.lab=0.4, cex.axis=0.4)
> axis(2)

 # predicted points will be visualized by color, actual points visualized by "A" or "S".
> colors <- c("blue", "red")
> colors <- colors[as.numeric(predicted.classes[,1])]
> letters <- c("S", "A")
> letters <- letters[as.numeric(ann)]

 # place points and legend on graph
> points(predicted.classes[,2],col=colors,pch=letters,cex=1.25)
> legend("bottomright", legend=c("Predicted Adeno", "Predicted Squamous"), fill=c("red", "blue"), cex = 0.5)
> legend("bottomleft", legend=c("Actual Squamous", "Actual Adeno"), pch=c("S", "A"), cex = 0.5)



8. Finally, using the top 5 discriminant genes (positive and negative direction) from your analysis, go to NCBI's DAVID (http://david.abcc.ncifcrf.gov/home.jsp) and look up the gene information. Provide the gene name and functional information (associated pathways, GO terms, etc) for these 10 genes.

 # find the fold change between the top genes from scc and adn
> fold <- apply(dat.2[,scc],1,mean) – apply(dat.2[,adn],1,mean)

 # sort fold change
> fold.sort <- sort(fold)

 # select the top 5 discriminat genes from negative and positive direction
 # negative
fold.sort[1:5]
219508_at 203717_at 205313_at
-2.465774 -2.335519 -2.108844
203914_x_at 243634_at
-1.882488 -1.611779

219508_at (glucosaminyl N-acetyl transferase 3, mucin type GCNT3)

219508_at	glucosaminyl (N-acetyl) transferase 3, mucin type(GCNT3)	Related Genes	Homo sapiens
GOTERM_BP_DIRECT	immunoglobulin production in mucosal tissue, carbohydrate metabolic process, protein O-linked glycosylation, O-glycan processing, tissue morphogenesis, intestinal absorption, kidney morphogenesis,		
GOTERM_CC_DIRECT	Golgi membrane, membrane, integral component of membrane, extracellular exosome,		
GOTERM_MF_DIRECT	beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase activity, N-acetyllactosaminide beta-1,6-N-acetylglucosaminyltransferase activity, acetylgalactosaminyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase activity,		
INTERPRO	Glycosyl transferase, family 14,		
KEGG_PATHWAY	Mucin type O-Glycan biosynthesis, Metabolic pathways,		
UP_KEYWORDS	Complete proteome, Disulfide bond, Glycoprotein, Glycosyltransferase, Golgi apparatus, Membrane, Proteomics identification, Reference proteome, Signal-anchor, Transferase, Transmembrane, Transmembrane helix,		
UP_SEQ_FEATURE	chain:Beta-1,3-galactosyl-O-glycosyl- glycoprotein beta-1,6-N- acetylglucosaminyltransferase 3, disulfide bond, glycosylation site:N-linked (GlcNAc...), topological domain:Cytoplasmic, topological domain:Lumenal, transmembrane region,		

203717_at (dipeptidyl peptidase 4 DPP4)

203717_at	dipeptidyl peptidase 4(DPP4)	Related Genes	Homo sapiens
GOTERM_BP_DIRECT	behavioral fear response, response to hypoxia, proteolysis, positive regulation of cell proliferation, negative regulation of extracellular matrix disassembly, T cell costimulation, regulation of cell-cell adhesion mediated by integrin, locomotory exploration behavior, psychomotor behavior, T cell activation, endothelial cell migration, viral entry into host cell,		
GOTERM_CC_DIRECT	lysosomal membrane, plasma membrane, focal adhesion, cell surface, membrane, integral component of membrane, apical plasma membrane, lamellipodium, endocytic vesicle, lamellipodium membrane, membrane raft, intercellular canaliculus, extracellular exosome, invadopodium membrane,		
GOTERM_MF_DIRECT	virus receptor activity, protease binding, serine-type endopeptidase activity, receptor binding, protein binding, serine-type peptidase activity, dipeptidyl-peptidase activity, identical protein binding, protein homodimerization activity,		
INTERPRO	Peptidase S9, prolyl oligopeptidase, catalytic domain, Peptidase S9B, dipeptidylpeptidase IV N-terminal, Peptidase S9, serine active site,		
KEGG_PATHWAY	Protein digestion and absorption,		
UP_KEYWORDS	3D-structure, Aminopeptidase, Cell adhesion, Cell junction, Cell membrane, Cell projection, Complete proteome, Direct protein sequencing, Disulfide bond, Glycoprotein, Hydrolase, Membrane, Protease, Proteomics identification, Receptor, Reference proteome, Secreted, Serine protease, Signal-anchor, Transmembrane, Transmembrane helix,		
UP_SEQ_FEATURE	active site:Charge relay system, chain:Dipeptidyl peptidase 4 membrane form, chain:Dipeptidyl peptidase 4 soluble form, disulfide bond, glycosylation site:N-linked (GlcNAc...), helix, mutagenesis site, sequence conflict, strand, topological domain:Cytoplasmic, topological domain:Extracellular, transmembrane region, turn,		

205313_at (HNF1 homeobox B HNF1B)

205313_at	HNF1 homeobox B(HNF1B)	Related Genes	Homo sapiens
GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter, endodermal cell fate specification, kidney development, inner cell mass cell differentiation, transcription, DNA-templated, Notch signaling pathway, response to glucose, anterior/posterior pattern specification, response to organic cyclic compound, insulin secretion, regulation of Wnt signaling pathway, hindbrain development, endocrine pancreas development, circadian regulation of gene expression, regulation of pronephros size, pronephric nephron tubule development, response to drug, regulation of endodermal cell fate specification, positive regulation of transcription, DNA-templated, positive regulation of transcription from RNA polymerase II promoter, embryonic digestive tract morphogenesis, branching morphogenesis of an epithelial tube, pronephros development, genitalia development, epithelial cell proliferation, positive regulation of transcription initiation from RNA polymerase II promoter, ureteric bud elongation, hepatoblast differentiation, negative regulation of mesenchymal cell apoptotic process involved in mesonephric nephron morphogenesis, protein-DNA complex assembly, hepatocyte differentiation, regulation of branch elongation involved in ureteric bud branching, mesonephric duct formation, negative regulation of mesenchymal cell apoptotic process involved in metanephros development,		
GOTERM_CC_DIRECT	nucleus, nucleoplasm, transcription factor complex,		
GOTERM_MF_DIRECT	RNA polymerase II regulatory region sequence-specific DNA binding, core promoter proximal region DNA binding, DNA binding, transcription factor activity, sequence-specific DNA binding, transcription factor activity, RNA polymerase II distal enhancer sequence-specific binding, protein binding, protein complex binding, protein homodimerization activity, sequence-specific DNA binding, protein heterodimerization activity,		
INTERPRO	Homeodomain, Hepatocyte nuclear factor 1, beta isoform, C-terminal, Hepatocyte nuclear factor 1, N-terminal, Homeodomain-like, Lambda repressor-like, DNA-binding domain, Hepatocyte nuclear factor 1, dimerisation domain,		
KEGG_PATHWAY	Maturity onset diabetes of the young,		
OMIM_DISEASE	Diabetes mellitus, noninsulin-dependent, Renal cysts and diabetes syndrome, Renal cell carcinoma,		
SMART	HOX,		
UP_KEYWORDS	3D-structure, Activator, Alternative splicing, Complete proteome, Diabetes mellitus, DNA-binding, Homeobox, Nucleus, Phosphoprotein, Polymorphism, Proteomics identification, Reference proteome, Transcription, Transcription regulation,		
UP_SEQ_FEATURE	chain:Hepatocyte nuclear factor 1-beta, DNA-binding region:Homeobox; HNF1-type, helix, modified residue, region of interest:Dimerization, sequence variant, splice variant, turn,		

203914_x_at (hydroxyprostaglandin dehydrogenase 15-NAD HPGD)

203914_x_at	hydroxyprostaglandin dehydrogenase 15-(NAD)(HPGD)	Related Genes	Homo sapiens
GOTERM_BP_DIRECT	prostaglandin metabolic process, transforming growth factor beta receptor signaling pathway, female pregnancy, parturition, lipoxygenase pathway, ovulation, negative regulation of cell cycle, oxidation-reduction process, thrombin receptor signaling pathway, ductus arteriosus closure, lipoxin metabolic process,		
GOTERM_CC_DIRECT	nucleoplasm, cytoplasm, cytosol, basolateral plasma membrane, extracellular exosome,		
GOTERM_MF_DIRECT	catalytic activity, prostaglandin E receptor activity, 15-hydroxyprostaglandin dehydrogenase (NAD+) activity, oxidoreductase activity, protein homodimerization activity, NAD binding, NAD+ binding,		
INTERPRO	Glucose/ribitol dehydrogenase, NAD(P)-binding domain, Short-chain dehydrogenase/reductase, conserved site,		
KEGG_PATHWAY	Transcriptional misregulation in cancer,		
OMIM_DISEASE	Digital clubbing, isolated congenital, Cranioosteoarthropathy, Hypertrophic osteoarthropathy, primary, autosomal recessive 1,		
UP_KEYWORDS	3D-structure, Alternative splicing, Complete proteome, Cytoplasm, Direct protein sequencing, Disease mutation, Fatty acid metabolism, Lipid metabolism, NAD, Oxidoreductase, Polymorphism, Prostaglandin metabolism, Proteomics identification, Reference proteome, Tumor suppressor,		
UP_SEQ_FEATURE	active site:Proton acceptor, binding site:NAD; via carbonyl oxygen, binding site:Substrate, chain:15-hydroxyprostaglandin dehydrogenase [NAD+], helix, mutagenesis site, nucleotide phosphate-binding region:NAD, sequence conflict, sequence variant, strand, turn,		

243634_at (small integral membrane protein 14 SMIM14)

243634_at	small integral membrane protein 14(SMIM14)	Related Genes	Homo sapiens
GOTERM_CC_DIRECT	endoplasmic reticulum, endoplasmic reticulum membrane, integral component of membrane,		
INTERPRO	Uncharacterised protein family, CD034/YQF4,		
UP_KEYWORDS	Complete proteome, Endoplasmic reticulum, Membrane, Proteomics identification, Reference proteome, Transmembrane, Transmembrane helix,		
UP_SEQ_FEATURE	chain:Uncharacterized protein C4orf34, transmembrane region,		

positive
> fold.sort[723:727]
204165_at 1562102_at 1552478_a_at
2.061723 2.167878 2.202305
209098_s_at 228038_at
2.842342 4.509514

204165_at (WAS protein family member 1 WASF1)

204165_at	WAS protein family member 1(WASF1)	Related Genes	Homo sapiens
BIOCARTA	Y branching of actin filaments, Rac 1 cell motility signaling pathway, How does salmonella hijack a cell,		
GOTERM_BP_DIRECT	protein complex assembly, movement of cell or subcellular component, Rac protein signal transduction, actin cytoskeleton organization, actin filament polymerization, lamellipodium morphogenesis, positive regulation of Arp2/3 complex-mediated actin nucleation,		
GOTERM_CC_DIRECT	mitochondrial outer membrane, cytoskeleton, focal adhesion, actin cytoskeleton, lamellipodium, SCAR complex, synapse,		
GOTERM_MF_DIRECT	actin binding, protein binding, protein complex binding, Rac GTPase binding,		
INTERPRO	WH2 domain,		
KEGG_PATHWAY	Adherens junction, Fc gamma R-mediated phagocytosis, Regulation of actin cytoskeleton, Bacterial invasion of epithelial cells, Shigellosis, Salmonella infection, Choline metabolism in cancer,		
SMART	WH2,		
UP_KEYWORDS	3D-structure, Actin-binding, Cell junction, Complete proteome, Cytoplasm, Cytoskeleton, Methylation, Phosphoprotein, Proteomics identification, Reference proteome, Synapse,		
UP_SEQ_FEATURE	chain:Wiskott-Aldrich syndrome protein family member 1, compositionally biased region:Poly-Pro, domain:WH2,		

1562102_at (not found on the DAVID website)

1552478_a_at (interferon regulatory factor 6 IRF6)

1552478_a_at	interferon regulatory factor 6(IRF6)	Related Genes	Homo sapiens
GOTERM_BP_DIRECT	transcription, DNA-templated, cell cycle arrest, negative regulation of cell proliferation, keratinocyte differentiation, keratinocyte proliferation, positive regulation of transcription, DNA-templated, cell development, interferon-gamma-mediated signaling pathway, type I interferon signaling pathway, mammary gland epithelial cell differentiation,		
GOTERM_CC_DIRECT	nucleus, cytoplasm, cytosol, extracellular exosome,		
GOTERM_MF_DIRECT	regulatory region DNA binding, DNA binding, transcription factor activity, sequence-specific DNA binding, protein binding,		
INTERPRO	Interferon regulatory factor DNA-binding domain, SMAD/FHA domain, Winged helix-turn-helix DNA-binding domain, SMAD domain-like, Interferon regulatory factor-3, Interferon regulatory factor, conserved site,		
OMIM_DISEASE	van der Woude syndrome, Popliteal pterygium syndrome 1, Orofacial cleft 6,		
SMART	IRF, SM01243,		
UP_KEYWORDS	Alternative splicing, Complete proteome, Cytoplasm, Differentiation, Disease mutation, DNA-binding, Nucleus, Polymorphism, Proteomics identification, Reference proteome, Transcription, Transcription regulation, Ubl conjugation,		
UP_SEQ_FEATURE	chain:Interferon regulatory factor 6, DNA-binding region:Tryptophan pentad repeat, sequence variant,		

209098_s_at (jagged 1 JAG1)

209098_s_at	jagged 1(JAG1)	Related Genes	Homo sapiens
BIOCARTA	Phosphoinositides and their downstream targets.,		
GOTERM_BP_DIRECT	angiogenesis, cell fate determination, blood vessel remodeling, morphogenesis of an epithelial sheet, T cell mediated immunity, pulmonary valve morphogenesis, cardiac right ventricle morphogenesis, Notch signaling pathway, Notch receptor processing, multicellular organism development, nervous system development, hemopoiesis, keratinocyte differentiation, regulation of cell migration, response to muramyl dipeptide, aorta morphogenesis, regulation of cell proliferation, auditory receptor cell differentiation, myoblast differentiation, endothelial cell differentiation, negative regulation of fat cell differentiation, positive regulation of myeloid cell differentiation, negative regulation of neuron differentiation, positive regulation of osteoblast differentiation, positive regulation of Notch signaling pathway, positive regulation of transcription from RNA polymerase II promoter, cardiac septum morphogenesis, ciliary body morphogenesis, pulmonary artery morphogenesis, cardiac neural crest cell development involved in outflow tract morphogenesis, Notch signaling involved in heart development, endocardial cushion cell development, nephron development, glomerular visceral epithelial cell development, distal tubule development, loop of Henle development, neuronal stem cell population maintenance, negative regulation of stem cell differentiation,		
GOTERM_CC_DIRECT	extracellular region, plasma membrane, integral component of plasma membrane, adherens junction, membrane, integral component of membrane, apical plasma membrane,		
GOTERM_MF_DIRECT	Notch binding, structural molecule activity, calcium ion binding, protein binding, growth factor activity,		
INTERPRO	EGF-type aspartate/asparagine hydroxylation site, Epidermal growth factor-like domain, von Willebrand factor, type C, Delta/Serrate/lag-2 (DSL) protein, EGF-like calcium-binding, Insulin-like growth factor binding protein, N-terminal, Notch ligand, N-terminal, EGF-like, conserved site, EGF, extracellular, EGF-like calcium-binding, conserved site, Jagged/Serrate protein,		
KEGG_PATHWAY	Notch signaling pathway, TNF signaling pathway,		
OMIM_DISEASE	Alagille syndrome, Tetralogy of Fallot, Deafness, congenital heart defects, and posterior embryotoxon,		
SMART	DSL, EGF_CA, EGF, VWC, VWC_out,		
UP_KEYWORDS	3D-structure, Alternative splicing, Calcium, Complete proteome, Developmental protein, Disease mutation, Disulfide bond, EGF-like domain, Glycoprotein, Membrane, Notch signaling pathway, Polymorphism, Proteomics identification, Reference proteome, Repeat, Signal, Transmembrane, Transmembrane helix,		
UP_SEQ_FEATURE	chain:Protein jagged-1, disulfide bond, domain:DSL, domain:EGF-like 10; calcium-binding, domain:EGF-like 11; calcium-binding, domain:EGF-like 12, domain:EGF-like 13, domain:EGF-like 14; calcium-binding, domain:EGF-like 15; calcium-binding, domain:EGF-like 1; atypical, domain:EGF-like 2, domain:EGF-like 3, domain:EGF-like 4; calcium-binding, domain:EGF-like 5; calcium-binding, domain:EGF-like 6; calcium-binding, domain:EGF-like 7; calcium-binding, domain:EGF-like 8, domain:EGF-like 9, glycosylation site:N-linked (GlcNAc...), sequence conflict, sequence variant, signal peptide, strand, topological domain:Cytoplasmic, topological domain:Extracellular, transmembrane region,		

228038_at (SRY-box SOX2)

228038_at	SRY-box 2(SOX2)	Related Genes	Homo sapiens
GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter, osteoblast differentiation, eye development, endodermal cell fate specification, chromatin organization, regulation of transcription, DNA-templated, transcription from RNA polymerase II promoter, cell cycle arrest, response to wounding, regulation of gene expression, glial cell fate commitment, pituitary gland development, adenohypophysis development, positive regulation of cell-cell adhesion, forebrain development, somatic stem cell population maintenance, tissue regeneration, regulation of cysteine-type endopeptidase activity involved in apoptotic process, positive regulation of MAPK cascade, positive regulation of cell differentiation, negative regulation of neuron differentiation, positive regulation of transcription, DNA-templated, positive regulation of transcription from RNA polymerase II promoter, inner ear development, negative regulation of epithelial cell proliferation, response to growth factor, negative regulation of canonical Wnt signaling pathway, neuronal stem cell population maintenance,		
GOTERM_CC_DIRECT	nucleus, nucleoplasm, transcription factor complex, cytoplasm, cytosol,		
GOTERM_MF_DIRECT	transcription regulatory region sequence-specific DNA binding, transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding, DNA binding, transcription factor activity, sequence-specific DNA binding, protein binding, miRNA binding, sequence-specific DNA binding, transcription regulatory region DNA binding,		
INTERPRO	High mobility group (HMG) box domain, Transcription factor SOX,		
KEGG_PATHWAY	Hippo signaling pathway, Signaling pathways regulating pluripotency of stem cells,		
OMIM_DISEASE	Microphthalmia, syndromic 3, Optic nerve hypoplasia and abnormalities of the central nervous system,		
SMART	HMG,		
UP_KEYWORDS	3D-structure, Activator, Complete proteome, Developmental protein, Disease mutation, DNA-binding, Isopeptide bond, Microphthalmia, Nucleus, Phosphoprotein, Reference proteome, Transcription, Transcription regulation, Ubl conjugation,		
UP_SEQ_FEATURE	chain:Transcription factor SOX-2, compositionally biased region:Poly-Ala, compositionally biased region:Poly-Gly, cross-link:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO), DNA-binding region:HMG box, helix,		